Robust Multi-view Learning via Half-quadratic Minimization

Yonghua Zhu[1,2], **Xiaofeng Zhu**[1,*], **Wei Zheng**[1]
[1] Guangxi Key Lab of Multi-source Information Mining & Security,
Guangxi Normal University, Guilin, 541004, China
[2] Guangxi University, Nanning, 530004, China
yhzhu66@qq.com, seanzhuxf@gmail.com, zwgxnu@163.com

Abstract

Although multi-view clustering is capable to use more information than single view clustering, existing multi-view clustering methods still have issues to be addressed, such as initialization sensitivity, the specification of the number of clusters, and the influence of outliers. In this paper, we propose a robust multi-view clustering method to address these issues. Specifically, we first propose a multi-view based sum-of-square error estimation to make the initialization easy and simple as well as use a sum-of-norm regularization to automatically learn the number of clusters according to data distribution. We further employ robust estimators constructed by the half-quadratic theory to avoid the influence of outliers for conducting robust estimations of both sum-of-square error and the number of clusters. Experimental results on both synthetic and real datasets demonstrate that our method outperforms the state-of-the-art methods.

1 Introduction

Compared to single-view clustering, multi-view clustering is capable to generate more robust clustering models as every view of multi-view data has particular information different from other views so that multi-view data provide more useful information for clustering [Chi and Lange, 2015; Zhang *et al.*, 2017; Nie *et al.*, 2017; Zhu *et al.*, 2014a]. Recently, reasonably taking full use of the information (*e.g.,* common information and particular information) among different views has been demonstrated to be a key strategy for multi-view clustering beyond single view clustering [Kriegel *et al.*, 2009; Hu *et al.*, 2017; Rodriguez and Laio, 2014].

By considering the way of using the information in multiple views, existing multi-view clustering methods can be categorized into three categories, such as concatenation approach, distributed approach, and centralized approach. Concatenation approach firstly concatenates the feature vectors of every view to form a set of long vectors and then conducts single view clustering on the resulting long-vector representations. Existing concatenation approach is easy to perform, but difficultly yields satisfied clustering performance. For example, high-dimensional representation of long vectors in concatenation approach easily leads to the issue of curse of dimensionality [Li *et al.*, 2017; Zhu *et al.*, 2016]. Distributed approach firstly generates clustering results from every view independently, and then proposes strategies to combine these results. Similar to the concatenation approach, distributed approach does not take full use of information of multi-view data yet and thus is unavailable to yield reasonable clustering performance [Sun, 2013]. Centralized approach takes advantage of distinguishing information or common information among multi-view data to more easily achieve better performance, compared to the former two approaches [Yin *et al.*, 2015]. However, centralized approach has well-known drawbacks such as initialization sensitivity [Shah and Koltun, 2017; Zhu *et al.*, 2018].

Increasingly complex datasets and real-world requirements are creating new demands for multi-view clustering [Xu *et al.*, 2016; Jiang *et al.*, 2015]. Real-world data are often corrupted by outliers [Zhu *et al.*, 2017b; Zhang *et al.*, 2006], but most existing multi-view clustering methods were designed to equivalently regard every sample without addressing the issue of the influence of outliers. Another challenging task of multi-view clustering is the specification of the number of clusters, which is required as an input. Usually, predefined input needs prior knowledge. Otherwise, the obtained solution is prone to error, especially for conducting multi-view clustering on the datasets with complex distribution or structure [Frey and Dueck, 2007; Fang and Wang, 2012].

In this paper, we propose a novel Robust Multi-View Clustering (RMVC) method to conduct simple initialization sensitivity, taking sample importance into account to avoid the influence of outliers, and automatically specifying the number of the clusters. To do this, firstly, a sum-of-square error estimation is used to minimize the difference between original samples and their corresponding new representations, aim for resulting in easy initialization, *i.e.,* initializing every new representation as its corresponding original feature to yield fast convergence. Moreover, we extend such a framework from single view data to multi-view data to further improve the clustering robustness by making the full use of the information in multi-view data. Meanwhile, we select robust estimators (based on the half-quadratic theory [Nikolova and Chan, 2007; He *et al.*, 2014]) to push a constraint on every sample of

[*] Corresponding author: Xiaofeng Zhu (seanzhuxf@gmail.com).

every view so that the outliers are removed out the model construction, unimportant samples are assigned small weights, and the important samples are assigned large weights. Secondly, a sum-of-norm regularization is used to control the tradeoff between the model fit (*i.e.,* the sum-of-square error estimation) and the number of clusters, so that the number of clusters can be automatically learnt according to the data distribution, and not need to be specified in advance.

Compared to previous multi-view clustering methods, our proposed RMVC method has the following advantages.

- Its initialization is easy as well as fixed, *i.e.,* initializing every sample with its original feature to achieve very simple and efficient initialization. It is noteworthy that existing multi-view clustering methods (*e.g.,* [Xu *et al.*, 2015; Nie *et al.*, 2017; Jiang *et al.*, 2015; Cao *et al.*, 2015]) are very sensitive to the initialization, and the initialization is complex such as a small set of random centroid is firstly initialized and then iteratively refined [Chi and Lange, 2015; Zhu *et al.*, 2014b]. Furthermore, different initializations lead to different clustering results. As a result, heuristic initialization methods should be used. Otherwise, clustering results may lead to a local minima and or even drastically suboptimal results.
- It does not require the input of the number of clusters, *i.e.,* automatically learning the number of clusters according to data distribution by a sum-of-norm regularization. By contrast, most of existing multi-view methods need to specify the number of clusters, which is time consuming as well as needs prior knowledge. Although some methods (*e.g.,* affinity propagation [Frey and Dueck, 2007; Zhu *et al.*, 2017a]) do not need to assign the number of clusters by users, these methods need to face the initialization issue [Shah and Koltun, 2017; Hocking *et al.*, 2011].
- It is capable of simultaneously separating outliers from important samples and learning the number of clusters. Most of existing multi-view methods ignored these two constraints and only a few of previous clustering methods considered one of them. Moreover, our method is a general methodology that can be applied to other clustering problems such as single view clustering and semi-supervised clustering.

2 Approach

We summarize all notations used in this paper in table 1.

2.1 Robust Clustering

K-means clustering partitions all the samples into K clusters so that every sample belongs to a specific cluster whose centroid is closest to this sample. Specifically, the objective function of k-means clustering [Ding *et al.*, 2005] can be formulated as:

$$\min_{\mathbf{G},\mathbf{F}} \|\mathbf{X}-\mathbf{G}\mathbf{F}\|_F^2 \quad s.t., G_{i,k}\in\{0,1\}, \sum_{k=1}^{K} G_{i,k}=1, i=1,...,n \tag{1}$$

$\mathbf{X}$	a matrix
$\mathbf{x}$	a vector of $\mathbf{X}$
$x_{i,j}$	the element in the *i*-th row and the *j*-th column of $\mathbf{X}$
$\mathbf{x}_{i,\cdot}$	the *i*-th row of $\mathbf{X}$
$\mathbf{x}_{\cdot,j}$	the *j*-th column of $\mathbf{X}$
$\|\|\mathbf{X}\|\|_F$	the Frobenius norm of $\mathbf{X}$, *i.e.,* $\|\|\mathbf{X}\|\|_F = \sqrt{\sum_{i,j}\mathbf{x}_{i,j}^2}$
$\|\|\mathbf{X}\|\|_{2,1}$	the $\ell_{2,1}$-norm of $\mathbf{X}$, *i.e.,* $\|\|\mathbf{X}\|\|_{2,1} = \sum_i \sqrt{\sum_j x_{ij}^2}$
$\mathbf{X}^T$	the transpose of $\mathbf{X}$
$tr(\mathbf{X})$	the trace of $\mathbf{X}$
$\mathbf{X}^{-1}$	the inverse of $\mathbf{X}$

Table 1: The detail of the notations used in this paper.

where $\mathbf{G}\in\mathbb{R}^{n\times k}$ is a binary matrix to indicate the samples belonging to a specific cluster out of K clusters, and $\mathbf{F}\in\mathbb{R}^{k\times d}$ is the centroid matrix, and $\|\cdot\|_F$ is a Frobenius norm.

Firstly, k-means clustering is sensitive to initialization since it is a non-convex optimization problem, whose optimization easily leads to a local minima or even sometimes drastically suboptimal. Moreover, heuristic initialization methods should be used for k-means clustering since different initializations may lead to considerably different clustering results [Hocking *et al.*, 2011]. Secondly, k-means clustering needs to set the number of clusters by users. It is usually time consuming for selecting an appropriate value for the number of clusters and there is a demand of prior knowledge. Thirdly, most existing clustering methods ignore the influence of outliers so that constructed clustering models deviate to the real ones.

Intuitively, closed samples in some similarity/distance measures should be assigned to the same cluster, and vice versa. As a result, a centroid can be obtained for each cluster. In real applications, we have no idea on the number of clusters, so we denote $\mathbf{u}_{i,\cdot}$ (the *i*-th row of $\mathbf{U}\in\mathbb{R}^{n\times d}$, $i=1,...,n$) as the new representation of the sample $\mathbf{x}_{i,\cdot}$ as well as the centroid containing the sample $\mathbf{x}_{i,\cdot}$. We expect to firstly initialize $\mathbf{u}_{i,\cdot}$ as $\mathbf{x}_{i,\cdot}$, and then recursively join the closed samples together to form a cluster/group until all the samples are jointed their closest clusters. This results in a sum-of-square error as follows:

$$\min_{\mathbf{U}} \sum_{i=1}^{n} \|\mathbf{x}_{i,\cdot}-\mathbf{u}_{i,\cdot}\|_2^2 \tag{2}$$

where $\|\cdot\|_2$ is an ℓ_2-norm. The minimization optimization of Eq. (2) easily results in trivial solutions, *i.e.,* $\mathbf{x}_{i,\cdot} = \mathbf{u}_{i,\cdot} (i=1,...,n)$. To address this issue, an appropriate constraint (*e.g.,* an ℓ_p-norm regularization) can be added to have the following formulation:

$$\min_{\mathbf{U}} \sum_{i=1}^{n} \|\mathbf{x}_{i,\cdot}-\mathbf{u}_{i,\cdot}\|_2^2 + \lambda \sum_{(p,q)\in\varepsilon} w_{p,q}\|\mathbf{u}_{p,\cdot}-\mathbf{u}_{q,\cdot}\|_p \tag{3}$$

where $\|\cdot\|_p$ is an ℓ_p-norm regularization, $(p,q)\in\varepsilon$ indicate that the *p*-th sample $\mathbf{x}_{p,\cdot}$ is one of the nearest neighbors of the *q*-th sample $\mathbf{x}_{q,\cdot}$, *i.e.,* their distance is less then ε, and λ is a tuning parameter balancing two terms of Eq. (3).

In Eq. (3), the first term (*i.e.,* the sum-of-square error estimation) learns the new representation $\mathbf{U}$ of $\mathbf{X}$ so that recursively updating $\mathbf{U}$ which is gradually close to $\mathbf{X}$ as well as satisfies the constraint in the second term, *i.e.,* the sum-of-norm regularization. The second term requires that $\mathbf{U}$ preserves the similarity of $\mathbf{X}$ and the new representations of samples from the same cluster are very approximate. Specifically, the similarity of $\mathbf{X}$ is preserved by an adjacency matrix $\mathbf{W} = [w_{p,q}]_{p,q=1,...,n} \in \mathbb{R}^{n \times n}$, which can be constructed by the k Nearest Neighbor (kNN) method, or the mutual kNN method [Brito *et al.*, 1997], or the sparse representation method [Elhamifar and Vidal, 2013]. On the other hand, if the *p*-th sample and the *q*-th sample come the same clustering, *i.e.,* $w_{p,q} \leq \varepsilon$, then their centroid should be same, *i.e.,* $\mathbf{u}_{p,\cdot} = \mathbf{u}_{q,\cdot}$. By inducing the ℓ_p-norm regularization and tuning the parameter λ, some same-cluster-pairs (*i.e.,* the pair $i \leftrightarrow j$ if $\mathbf{u}_{i,\cdot} = \mathbf{u}_{j,\cdot}, i, j = 1, ...n$) are generated. Moreover, the larger the value of λ, the more the number of same-cluster-pairs is. After the optimization process, all the samples are gathered into clusters, and thus the number of the cluster is decided.

Eq. (3) solves the issue of initialization sensitivity, but does not solve the issue of the specification of the number of clusters well as the ℓ_p-norm regularization has still limitations on robustly dealing with the outlier edges of $\mathbf{W}$, high-order outliers for short. That is, the outlier edges will still appear instead of diminishing during the process of optimizing $\mathbf{U}$. Moreover, Eq. (3) does not touch the issue of outlier samples (low-order outliers for short) in its first term. In this paper, we have two considerations. Firstly, the outlier edges in $\mathbf{W}$ will be suppressed via assigning a small or even zero weight on the difference estimation of $\|\mathbf{u}_{p,\cdot} - \mathbf{u}_{q,\cdot}\|$. Secondly, the outliers in $\mathbf{X}$ will also be assigned small weights. To do this, we employ robust estimators constructed by the half-quadratic theory to have:

$$\min_{\mathbf{U}} \sum_{i=1}^{n} \rho_1(\|\mathbf{x}_{i,\cdot} - \mathbf{u}_{i,\cdot}\|_2) + \lambda \sum_{(p,q)\in\varepsilon} w_{p,q}\rho_2(\|\mathbf{u}_{p,\cdot} - \mathbf{u}_{q,\cdot}\|_2) \tag{4}$$

where $\rho_1(\cdot)$ and $\rho_2(\cdot)$ are robust estimators constructed by the half-quadratic theory [Nikolova and Ng, 2005; Charbonnier *et al.*, 1997; Black and Rangarajan, 1996]. Robust estimators were designed to produce robust estimation that are not unduly affected by outliers and output good performance when there are small departures from parametric distributions [Nikolova and Ng, 2005]. The popular robust estimators include $\ell_1 - \ell_2$ estimator, Cauchy estimator, Geman-McClure estimator, and so on.

2.2 Robust Multi-view Clustering

We denote $\mathbf{X}^v \in \mathbb{R}^{n \times d^v}$ of multi-view data as the feature matrix of the *v*-th view data, where *n* and d^v, respectively, are the number of samples and the features of the *v*-th view data. We also denote $x^v_{i,j}, \mathbf{x}^v_{i,\cdot}$ and $\mathbf{x}^v_{\cdot,j}$, respectively, as the element of the *i*-th row and the j-th column, the vector of the *i*-th row, and the vector of the j-th column, of the *v*-th feature matrix. It is noteworthy that the dimensions of different views are usually different. In this paper, we first map the feature matrix $\mathbf{X}^v$ of each view into a kernel space $\hat{\mathbf{X}}^v \in \mathbb{R}^{n \times n}$ via a linear kernel. Thus, we extend Eq. (4) into the following Robust Multi-View Clustering (RMVC):

$$\begin{aligned} \min_{\hat{\mathbf{U}}} & \sum_{v=1}^{m}\sum_{i=1}^{n} \rho_1(\|\hat{\mathbf{x}}^v_{i,\cdot} - \hat{\mathbf{u}}_{i,\cdot}\|_2) \\ & + \lambda \sum_{v=1}^{m} \sum_{(p,q)\in\varepsilon} \hat{w}^v_{p,q}\rho_2(\|\hat{\mathbf{u}}_{p,\cdot} - \hat{\mathbf{u}}_{q,\cdot}\|_2) \end{aligned} \tag{5}$$

In Eq. (5), we expect to learn a common representation $\hat{\mathbf{U}} \in \mathbb{R}^{n \times n}$ from all *m* views and save the similarity of every view in $\hat{\mathbf{W}}^v$ $(v = 1, ..., n)$ which are constructed by the mutual kNN method in this paper. After optimizing Eq. (5), we use the obtained $\hat{\mathbf{U}}$ (*i.e.,* the common representation of multi-view data) to conduct multi-view clustering.

For simplicity, in this paper, we use the $\ell_1 - \ell_2$ estimator to replace both $\rho_1(\cdot)$ and $\rho_2(\cdot)$, *i.e.,*

$$\begin{cases} \rho_1(\mathbf{z}) = \sqrt{\gamma_1 + \mathbf{z}^2} \\ \rho_2(\mathbf{z}) = \sqrt{\gamma_2 + \mathbf{z}^2} \end{cases} \tag{6}$$

where $\mathbf{z}$ is a vector or matrix variable, γ_1 and γ_2 are used to control the number of samples involving into the process of optimization in each iteration. Hence, we obtain the final objective function as follows:

$$\begin{aligned} \min_{\hat{\mathbf{U}}} & \sum_{v=1}^{m}\sum_{i=1}^{n} \sqrt{\gamma^v_1 + \|\hat{\mathbf{x}}^v_{i,\cdot} - \hat{\mathbf{u}}_{i,\cdot}\|_2^2} \\ & + \lambda \sum_{v=1}^{m} \sum_{(p,q)\in\varepsilon} \hat{w}^v_{p,q}\sqrt{\gamma_2 + \|\hat{\mathbf{u}}_{p,\cdot} - \hat{\mathbf{u}}_{q,\cdot}\|_2^2} \end{aligned} \tag{7}$$

where γ^v_1 is the tuning parameter of the *v*-th view data $\hat{\mathbf{X}}^v$.

For efficient optimization and the control of the number of samples involving into the process of optimization in each iteration, the half-quadratic theory usually transfers the robust estimator into its equivalent formulation via adding an extra variable $\mathbf{c}$, *i.e.,* $\min_{\mathbf{z}} \rho(\mathbf{z}) \Longleftrightarrow \min_{\mathbf{c},\mathbf{z}} \mathbf{c}\mathbf{z}^2 + \psi(\mathbf{c})$, where $\mathbf{c}\mathbf{z}^2$ is a quadratic function, $\psi(\cdot)$ is the dual function of the robust estimator [He *et al.*, 2014]. Usually, different robust estimators have different dual functions and every dual function has its minimization function which decides the weight vector $\mathbf{c}$. After such a transformation, each element c_i in the weight vector $\mathbf{c}$ is the weight of the corresponding z_i. Specifically, if z_i is an outlier, then the value of c_i is small or even 0, and vice versa. Hence, the influence of outliers is suppressed or even diminished while tuning the parameters of robust estimators, such as γ^v_1 and γ_2 in Eq. (8). Moreover, we do not need to know the explicit function of the dual function as $\mathbf{c}$ is decided by the minimization function of the dual function.

Based on above analysis, we transfer our objective function in Eq. (7) to the following problem:

$$\begin{aligned} \min_{\hat{\mathbf{U}},\alpha^v,\beta} & \sum_{v=1}^{m} (\sum_{i=1}^{n} \alpha^v_i \|\hat{\mathbf{x}}^v_{i,\cdot} - \hat{\mathbf{u}}_{i,\cdot}\|_2^2 + \psi_1(\alpha^v)) \\ & + \lambda \sum_{v=1}^{m} \sum_{(p,q)\in\varepsilon} \hat{w}^v_{p,q}\beta_{p,q}\|\hat{\mathbf{u}}_{p,\cdot} - \hat{\mathbf{u}}_{q,\cdot}\|_2^2 + \psi_2(\beta) \end{aligned} \tag{8}$$

where $\alpha^v \in \mathbb{R}^{n \times 1}$ and $\beta \in \mathbb{R}^{n \times n}$. The vector α^v is used to control outlier samples in $\mathbf{X}^v$ and β is used to control the outlier edges in $\hat{\mathbf{W}}^v$.

Eq. (8) is not convex for all the variables (*i.e.,* $\mathbf{U}$, α^v $(v = 1, ..., m)$, β) but is convex for each of them while fixing the others. In this paper, we employ the alternative optimization strategy to optimize Eq. (8), *i.e.,* (i) optimizing α^v and β while fixing $\mathbf{U}$ and (ii) optimizing $\mathbf{U}$ while fixing α^v and β. We literately repeat the step (i) and (ii) until the objective value achieves stable or satisfies the predefined demand of stop criteria.

2.3 Convergence

To prove the convergence of optimization algorithm (*i.e.,* the alternative strategy), we first denote $\hat{\mathbf{U}}^{(t+1)}$ as the value of (t +1)-th iteration of $\hat{\mathbf{U}}$. Thus Eq. (8) is changed into:

$$\begin{aligned} &J(\alpha^{v(t+1)}, \beta^{(t+1)}, \hat{\mathbf{U}}^{(t+1)}) = \\ &\quad \sum_{v=1}^{m}(\sum_{i=1}^{n} \alpha_i^{v(t+1)} \|\hat{\mathbf{x}}_{i,\cdot}^{v} - \hat{\mathbf{u}}_{i,\cdot}^{(t+1)}\|_2^2 + \psi_1(\alpha^{v(t+1)}))+ \\ &\lambda \sum_{v=1}^{m} \sum_{(p,q)\in\varepsilon} \hat{w}_{p,q}^{v} \beta_{p,q}^{(t+1)} \|\hat{\mathbf{u}}_{p,\cdot}^{(t+1)} - \hat{\mathbf{u}}_{q,\cdot}^{(t+1)}\|_2^2 + \psi_2(\beta^{(t+1)}) \end{aligned} \tag{9}$$

As the variables α^v and β are decided by the half-quadratic theory, whose convergence has been proved in [Nikolova and Ng, 2005]. Thus we have

$$J(\alpha^{v(t+1)}, \beta^{(t+1)}, \hat{\mathbf{U}}^{(t+1)}) \leq J(\alpha^{v(t)}, \beta^{(t)}, \hat{\mathbf{U}}^{(t+1)}) \tag{10}$$

The variable $\hat{\mathbf{U}}$ has a closed form solution, so

$$J(\alpha^{v(t)}, \beta^{(t)}, \hat{\mathbf{U}}^{(t+1)}) \leq J(\alpha^{v(t)}, \beta^{(t)}, \hat{\mathbf{U}}^{(t)}) \tag{11}$$

By plugging Eq. (11) into Eq. (10), we have:

$$J(\alpha^{v(t+1)}, \beta^{(t+1)}, \hat{\mathbf{U}}^{(t+1)}) \leq J(\alpha^{v(t)}, \beta^{(t)}, \hat{\mathbf{U}}^{(t)}) \tag{12}$$

Hence, optimization algorithm converges to a local minima.

3 Experiments

We evaluated our proposed method by comparing with three state-of-the-art multi-view clustering and a single view clustering method on one synthetic multi-view dataset and four real multi-view datasets in term of four clustering metrics, such as ACCuracy of clustering (ACC), Normalized Mutual Information (NMI), Purity, and Adjusted Rand Index (ARI).

3.1 Experimental Settings

In our experiments, we evaluated our method by comparing with five comparison methods, which details were summarized as follows:

- We conducted k-means clustering on every view of multi-view data and denoted the worst result and best result, respectively, as **Worst** and **Best**.
- As a concatenation approach, Concatenation k-means clustering (**ConKM**) concatenates the features cross all the views to form a long vector and then conducts k-means clustering on a single dataset.
- As a distributed approach, Weighted view Collaborative Fuzzy C-means (**WvCoFCM**) [Jiang *et al.*, 2015] first conducts clustering on each view based on the fuzzy k-means clustering and then combines all the resulting result to output the final clustering result. WvCoFCM takes the importance of each view into account.
- As a centralized approach, Multi-View Self-Paced Learning for Clustering (**MSPLC**) [Xu *et al.*, 2015] uses self-paced learning to simultaneously consider both the sample-diversity and the view-diversity.

Data Sets	#(Views)	#(Samples)	#(Classes)	#(Type)
Cornell	4	195	5	Web
Texas	4	187	5	Web
Washingdon	4	230	5	Web
Wisconsin	4	265	5	Web

Table 2: The details of public multi-view datasets.

In our experiments, we repeated k-means clustering 20 times and reported their average value for single view clustering and ConKM, and set the value of k as $[10, 15, ..., 40]$ in k-nearest neighbor for the methods such as our proposed method. We also set the number of clusters as the number of real classes for all comparison methods, while our method automatically generates the number of clusters. The ranges of parameters of every method were set by strictly following the corresponding literature. We set the range of parameter λ in our method as $\lambda \in [10^{-3}, 10^{-2}, ..., 10^{3}]$ and set the stop criteria of the iterative optimization as $\frac{\|obj(t+1)-obj(t)\|_2^2}{obj(t)} \leq 10^{-6}$, where $obj(t)$ stands for the objective value in the t-th iteration.

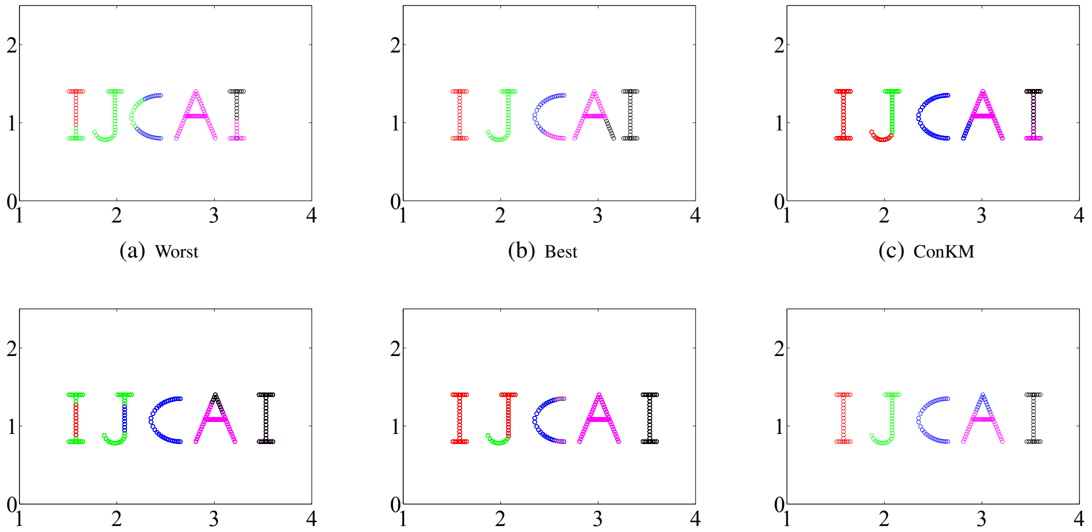
3.2 Synthetic Datasets

The synthetic dataset consisted of five letters (such as "I", "J", "C", "A", and "I"), which were represented by five views. The samples in each letter formed a cluster and were visualized with a color different from others. The first letter and the last letter were far away other letters in first two views, the former two letters and the last two letters were far away from other letters in other two views, and the middle letter was far away from other letters in 5th view. Figure 1 visualized the clustering results of all the methods, where all the clustering methods did not correctly output exact clustering results since the shapes of clusters are different and irregular. However, our method achieved the best clustering performance, compared to all the comparison methods. Actually, it is difficult for the comparison methods to have effective initializations on complex multi-view datasets. For example, random sampling for the initialization was unavailable to obtain initial centroid cross all the clusters, so the final clustering results easily achieved local minima or even sub-optimality. By contrast, our proposed method exactly selected all centroid since the initialization with original samples preserved the structures of the data better than random sampling methods. Thus, our method could easily achieve good clustering performance with simple initialization.

In a word, our method outperformed all the comparison methods for multi-view clustering on the synthetic dataset, which was with predefined number of clusters. The experimental results verified the advantages of our method in term of initialization sensitivity via the synthetic dataset.

3.3 Real Datasets

We used four real multi-view datasets to evaluate the practical clustering performance of our proposed method. The details of the used datasets are reported in Table 2 and the clustering



(a) Worst (b) Best (c) ConKM

(d) WvCoFCM (e) MSPLC (f) Proposed

Figure 1: Cluster results of the synthetic dataset.

results of all the methods on these datasets were reported in Table 3.

Obviously, our method achieved the best clustering performance, followed by WvCoFCM, MSPLC and ConKM. For example, the accuracy of our method improved on average by 8.71% and 12.88%, respectively, compared to the best comparison method (*i.e.,* WvCoFCM) and the worst comparison method (*i.e.,* ConKM) on all four datasets. The possible reason is that our method simultaneously takes three constraints into account, *i.e.,* initialization sensitivity, the influences of outliers, and the specification of the number of clusters. By contrast, all the comparison methods only consider a part of these constraints, so that were unavailable to output robust clustering models. In particular, both our proposed method and MSPLC considered the influence of outliers, but MSPLC does not touch other constraints and only took the influence of the low-order outliers into account.

3.4 Parameters' Tuning

In this section, we investigated the variations of our method with different parameters' setting.

Firstly, we conducted experiments on our method with different parameters (k and λ) and listed the ACC results in Figure 2. The results showed that our method is sensitive a little to parameters' setting. For example, our method easily achieved the best clustering performance while setting $k \in [10, 20]$ and $\lambda \in [10^1, 10^3]$ on dataset Wisconsin. This indicates that the parameters ranges of our method can be easily adjusted to achieve reasonable clustering performance.

Secondly, we experimentally demonstrated the convergence of our proposed optimization algorithm to solve our objective function Eq. (7) via listing the variations of objective function values in Eq. (7) with different numbers of iterations in Figure 3. As a result, our proposed optimization algorithm fast converged within 20 iterations.

Thirdly, we conducted experiments to show the feasibility of automatically generating the number of clusters. To do this, we generated different binary matrices via tuning the parameters in our method so that our method outputted different numbers of clusters. We reported the results in Figure 4, where yellow color is the results while the number of clusters is the real classes. From Figure 4, the datasets partitioned into different numbers of clusters may yield different clustering results. Moreover, the yellow results may be worse than other results. This indicates that the real class may not be the best number of clusters. This is true in real applications because one cluster may be partitioned into multiple sub-clusters and some clusters may also be categorized into one cluster [Zhu and Martinez, 2006]. Moreover, the results in Figure 4 verified the advantages of our proposed method in terms of the specification of the number of clusters, *i.e.,* flexibly outputting the number of clusters according to data distribution and not needing to preset the number of clusters for clustering.

4 Conclusion

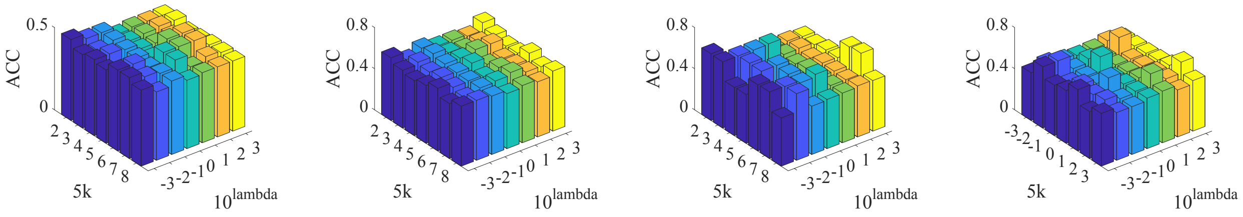
In this paper, we have proposed a novel robust multi-view clustering method to deal with the issues of existing multi-view clustering, such as initialization sensitivity, the influence of outliers, and the specification of the number of clusters. Experimental results on both synthetic and real datasets verified that these constraints affect clustering results as our method considering all of them outperformed the comparison methods only considering a part of these constraints.

Acknowledgements

This work was supported in part by the Nation Natural Science Foundation of China (Grants No: 61573270 and 61672177), the China Key Research Program (Grant No: 2016YFB1000905), the Guangxi Natural Science Foundation

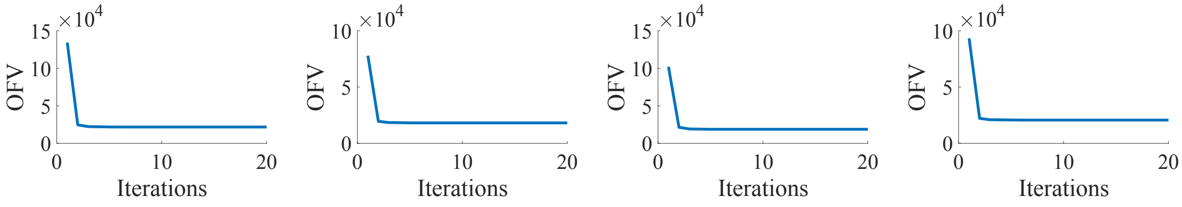
Datasets	Cornell				Texas				Washingdon				Wisconsin			
	ACC	NMI	Purity	ARI	ACC	NMI	Purity	ARI	ACC	NMI	Purity	ARI	ACC	NMI	Purity	ARI
Worst	0.3641	0.0578	0.4308	0.0394	0.4439	0.2086	0.6043	0.0768	0.4478	0.0278	0.4652	0.0084	0.4113	0.0328	0.4679	0.0230
Best	0.4256	0.0274	0.4359	0.0081	0.5561	0.1506	0.5829	0.1673	0.4739	0.0259	0.4826	0.0076	0.5208	0.1614	0.5396	0.1808
ConKM	0.4359	0.2709	0.5744	0.1336	0.4492	0.2048	0.6086	0.1297	0.4565	0.0148	0.5696	0.0057	0.4906	0.0496	0.4981	0.0181
WVCoFCM	0.4513	0.3032	0.5897	**0.1499**	0.5027	**0.3382**	0.6952	0.2287	0.4913	0.2925	0.6565	0.2016	0.5509	0.1827	0.5547	**0.1889**
MSPLC	0.3995	0.2288	0.4462	0.0167	0.4385	0.2568	0.5615	0.0271	0.4774	0.3216	0.6470	0.2508	0.5102	0.3445	0.5509	0.0779
Proposed	**0.4945**	**0.3125**	**0.6006**	0.1482	**0.6087**	0.3215	**0.7067**	**0.2378**	**0.6478**	**0.3642**	**0.7054**	**0.3379**	**0.5964**	**0.3538**	**0.6304**	0.1791

Table 3: Clustering results on real multi-view data sets.



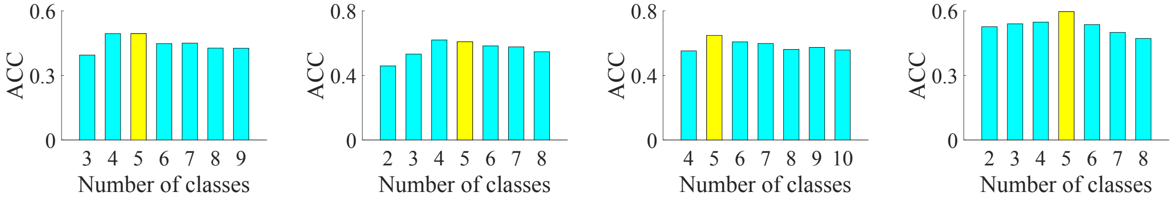
(a) Cornell (b) Texas (c) Washingdon (d) Wisconsin

Figure 2: The variation of different parameter combination.



(a) Cornell (b) Texas (c) Washingdon (d) Wisconsin

Figure 3: The Objective Function Value (OFV) of our method with different iterations.



(a) Cornell (b) Texas (c) Washingdon (d) Wisconsin

Figure 4: ACC of our method with different numbers of clusters, where yellow-color represented the clustering results with the real classes.

(Grant No: 2015GXNSFCB139011), the Guangxi High Institutions Program of Introducing 100 High-Level Overseas Talents, the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing, and the Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (18-A-01-01).

References

[Black and Rangarajan, 1996] Michael J. Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19(1):57–91, 1996.

[Brito *et al.*, 1997] MR Brito, EL Chavez, AJ Quiroz, and JE Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics & Probability Letters*, 35(1):33–42, 1997.

[Cao *et al.*, 2015] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, pages 586–594, 2015.

[Charbonnier *et al.*, 1997] Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6(2):298–311, 1997.

[Chi and Lange, 2015] Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

[Ding *et al.*, 2005] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDMs*, pages 606–610, 2005.

[Elhamifar and Vidal, 2013] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.

[Fang and Wang, 2012] Yixin Fang and Junhui Wang. Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 56(3):468–477, 2012.

[Frey and Dueck, 2007] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[He *et al.*, 2014] Ran He, Wei-Shi Zheng, Tieniu Tan, and Zhenan Sun. Half-quadratic-based iterative minimization for robust sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):261–275, 2014.

[Hocking *et al.*, 2011] Toby Dylan Hocking, Armand Joulin, Francis Bach, and Jean-Philippe Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *ICML*, page 1, 2011.

[Hu *et al.*, 2017] Rongyao Hu, Xiaofeng Zhu, Debo Cheng, Wei He, Yan Yan, Jingkuan Song, and Shichao Zhang. Graph self-representation method for unsupervised feature selection. *Neurocomputing*, 220:130–137, 2017.

[Jiang *et al.*, 2015] Yizhang Jiang, Fu-Lai Chung, Shitong Wang, Zhaohong Deng, Jun Wang, and Pengjiang Qian. Collaborative fuzzy clustering from multiple weighted views. *IEEE Trans. Cybernetics*, 45(4):688–701, 2015.

[Kriegel *et al.*, 2009] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *TKDD*, 3(1):1:1–1:58, 2009.

[Li *et al.*, 2017] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6):94:1–94:45, 2017.

[Nie *et al.*, 2017] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, pages 2408–2414, 2017.

[Nikolova and Chan, 2007] Mila Nikolova and Raymond H. Chan. The equivalence of half-quadratic minimization and the gradient linearization iteration. *IEEE Trans. Image Processing*, 16(6):1623–1627, 2007.

[Nikolova and Ng, 2005] Mila Nikolova and Michael K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Scientific Computing*, 27(3):937–966, 2005.

[Rodriguez and Laio, 2014] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[Shah and Koltun, 2017] Sohil Atul Shah and Vladlen Koltun. Robust continuous clustering. *Proceedings of the National Academy of Sciences of the United States of America*, 114(37):9814, 2017.

[Sun, 2013] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

[Xu *et al.*, 2015] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view self-paced learning for clustering. In *IJCAI*, pages 3974–3980, 2015.

[Xu *et al.*, 2016] Jinglin Xu, Junwei Han, Kai Xiong, and Feiping Nie. Robust and sparse fuzzy k-means clustering. In *IJCAI*, pages 2224–2230, 2016.

[Yin *et al.*, 2015] Qiyue Yin, Shu Wu, Ran He, and Liang Wang. Multi-view clustering via pairwise sparse subspace representation. *Neurocomputing*, 156:12–21, 2015.

[Zhang *et al.*, 2006] Shichao Zhang, Yongsong Qin, Xiaofeng Zhu, Jilian Zhang, and Chengqi Zhang. Optimized parameters for missing data imputation. In *PRICAI*, pages 1010–1016, 2006.

[Zhang *et al.*, 2017] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. Efficient knn classification with different numbers of nearest neighbors. *IEEE transactions on neural networks and learning systems,DOI: 10.1109/TNNLS.2017.2673241*, 2017.

[Zhu and Martinez, 2006] Manli Zhu and Aleix M Martinez. Subclass discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1274–1286, 2006.

[Zhu *et al.*, 2014a] Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen. Multi-modality canonical feature selection for alzheimer's disease diagnosis. In *MICCAI*, pages 162–169, 2014.

[Zhu *et al.*, 2014b] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *CVPR*, pages 1542–1549, 2014.

[Zhu *et al.*, 2016] Yingying Zhu, Xiaofeng Zhu, Minjeong Kim, Dinggang Shen, and Guorong Wu. Early diagnosis of alzheimer's disease by joint feature selection and classification on temporally structured support vector machine. In *MICCAI*, pages 264–272, 2016.

[Zhu *et al.*, 2017a] Pengfei Zhu, Wencheng Zhu, Qinghua Hu, Changqing Zhang, and Wangmeng Zuo. Subspace clustering guided unsupervised feature selection. *Pattern Recognition*, 66:364–374, 2017.

[Zhu *et al.*, 2017b] Yingying Zhu, Xiaofeng Zhu, Minjeong Kim, Daniel Kaufer, and Guorong Wu. A novel dynamic hyper-graph inference framework for computer assisted diagnosis of neuro-diseases. In *IPMI*, pages 158–169. Springer, 2017.

[Zhu *et al.*, 2018] Pengfei Zhu, Qian Xu, Qinghua Hu, Changqing Zhang, and Hong Zhao. Multi-label feature selection with missing labels. *Pattern Recognition*, 74:488–502, 2018.